

Inferring the origin of an epidemic with dynamic message-passing algorithm

Andrey Y. Lokhov¹, Marc Mézard^{1,2}, Hiroki Ohta¹, and Lenka Zdeborová³

¹Université Paris-Sud/CNRS, LPTMS, UMR8626, Bât. 100, 91405 Orsay, France,

²Ecole normale supérieure, 45 rue d'Ulm, 75005 Paris, France, and

³Institut de Physique Théorique, IPhT, CEA Saclay and CNRS URA 2306, 91191 Gif-sur-Yvette, France.

(Dated: March 22, 2013)

We study the problem of estimating the origin of an epidemic outbreak – given a contact network and a snapshot of epidemic spread at a certain time, determine the infection source. Finding the source is important in different contexts of computer or social networks. We assume that the epidemic spread follows the most commonly used susceptible-infected-recovered model. We introduce an inference algorithm based on dynamic message-passing equations, and we show that it leads to significant improvement of performance compared to existing approaches. Importantly, this algorithm remains efficient in the case where one knows the state of only a fraction of nodes.

PACS numbers: 89.75.Hc, 05.20.-y, 02.50.Tt

Introduction: Understanding and controlling the spread of epidemics on networks of contacts is an important task of today's science. It has far-reaching applications in mitigating the results of epidemics caused by infectious diseases, computer viruses, rumor spreading in social media and others. In the present article we address the problem of estimation of the origin of the epidemic outbreak (the so-called patient zero, or infection source – in what follows, these terms are used interchangeably): given a contact network and a snapshot of epidemic spread at a certain time, determine the infection source. Information about the origin could be extremely useful to reduce or prevent future outbreaks. Whereas the dynamics and prediction of epidemic spreading in networks has attracted a considerable number of works, for a review see [1–3], the problem of estimation of the epidemic origin has been mathematically formulated only recently [4], followed by a burst of research on this practically important problem [5–11]. In order to make the estimation of the origin of spreading a well defined problem we need to have some knowledge about the spreading mechanism. We shall adopt here the same framework as in existing works, namely we assume that the epidemic spread follows the widely used susceptible-infected-recovered (SIR) model or some of its special cases [12].

The stochastic nature of infection propagation makes the estimation of epidemic origin intrinsically hard: indeed, different initial conditions can lead to the same configuration at the observation time. Finding an estimator that maximizes the probability of the observed configuration is in general computationally intractable, except in very special cases such as the case where the contact network is a line or a regular tree [4, 6, 11]. The methods that have been studied in the existing works are mostly based on various kinds of graph centrality measures. Examples include the distance centrality or the Jordan center of a graph [4–7]. The problem was generalized to estimation of a set of epidemic origins using spectral methods in [8, 9]. Another line of approach uses more involved information about the epidemic than a snapshot at a given time [10].

In this paper we introduce a new algorithm for the estimation of the origin of an SIR epidemic from the knowledge of the network and the snapshot of some nodes at a certain time. Our algorithm estimates the probability that the observed snapshot resulted from a given patient zero in a way which is crucially different from existing approaches. For every possible origin of the epidemic, we use a fast dynamic message-passing method to estimate the probability that a given node in the network was in the observed state (S , I or R). We then use a mean-field-like approximation to compute the probability of the observed snapshot as a product of the marginal probabilities. We finally rank the possible origins according to that probability.

The dynamic message-passing (DMP) algorithm that we use to estimate the probability of a given node to be in a given state is interesting in itself. It is based on dynamic equations that were first suggested in a different (and not straightforwardly tractable) form in [13]. If averaged over a graph ensemble it leads to the asymptotically exact dynamic equations of [14, 15] for SIR, or to those of [16] for avalanches in the random field Ising model. Note, that DMP, although it bears some similarity with the standard belief propagation (BP) method [17, 18], is crucially different from BP since it does not derive from a Boltzmann-like probability distribution. It does not need to be iterated till convergence, instead the iteration time corresponds directly to the real time in the associated SIR dynamics. A nice property that DMP shares with BP is that it is exact if the contact network is a tree. We use it as an approximation for loopy but sparse contact networks in the same way that BP is commonly used with success in such situations.

We test our algorithm on synthetic spreading data and show that it performs better than existing approaches (except for a special region of parameters where the Jordan center is on average better). We find that the algorithm is very robust, for instance it remains efficient even in the case when the state of only a fraction of nodes in the network is observed. From our tests we also identify regions of parameters where estimating the origin of

epidemic spreading is relatively easy and others where this problem hard. Our dataset can hence also serve as a test-bed for new approaches.

SIR spreading model and inference of epidemic origin:

Let $G \equiv (V, E)$ be a connected undirected graph containing N nodes defined by the set of vertices V and the set of edges E . The SIR model for spreading of an epidemic is defined as follows: Each node i at discrete time t can be in one of three states $q_i(t)$: susceptible, $q_i(t) = S$, infected, $q_i(t) = I$, or recovered, $q_i(t) = R$. At each time step, an infected node will recover with probability μ_i , and a susceptible node i will become infected with probability $1 - \prod_{k \in \partial i} (1 - \lambda_{ki} \delta_{q_k(t), I})$, where ∂i is the set of the nodes neighboring node i , and λ_{ki} measures the efficiency of spread from k to i . The recovered nodes never change their state. We assume that the graph G and parameters λ_{ij} , μ_i are known (or have been inferred). The general properties and the phase diagram of this model on random networks were studied in many works, see e.g. [3] and references therein.

To define the problem of estimation of the epidemic origin we consider only the case where at initial time only one node is infected (this node will be referred to as the patient zero, i_0), and all others nodes are initially susceptible. After $t_0 > 0$ time steps (t_0 is in general unknown), we observe the state of a set of nodes $\mathcal{O} \subset V$, and the task is to estimate the location of patient zero based on this snapshot.

Let us briefly explain two of existing algorithms [4, 6, 7], that we will use as benchmarks. The authors of [4, 6, 7] considered only the case when all the nodes were observed, $\mathcal{O} = V$. A version of these algorithms for the more general case is suggested in appendix B. The most basic measure for node i to be the epidemic origin is the distance centrality $D(i)$ which we define as $D(i) \equiv \sum_{j \in \mathcal{G}} d(i, j) (\delta_{q_j, I} + \delta_{q_j, R} / \mu_j)$, where the graph \mathcal{G} is a connected component of the original graph G containing all infected and recovered nodes and only them, and $d(i, j)$ is the shortest path between node i and node j on the graph \mathcal{G} . The ad-hoc factor $1/\mu_j$ is introduced to distinguish recovered nodes that for small μ_j tend to be closer to the epidemic origin. In the existing works this factor was not present, because [4, 6] treated only the SI model, and [7] considered that susceptible and recovered nodes are indistinguishable. The authors of [4, 6] suggested a ‘‘rumor centrality’’ estimator and showed that, for tree graphs, the rumor centrality and the distance centrality coincide. Another simple but well performing estimator, Jordan centrality $J(i)$, was proposed in [7] and corresponds to a node minimizing the maximum distance to other infected and recovered nodes: $J(i) \equiv \max_{j \in \mathcal{G}} d(i, j)$. This estimator is known as Jordan center of \mathcal{G} in the graph theory literature.

The core of the algorithm proposed in the present work is dynamic message-passing, explained in the next section, that is able to estimate for a given patient zero and a given observation time t_0 what is the probability that a node i will be observed in a given state. For simplicity of

the explanation let us first assume the time t_0 is known. Let us call $P_S^j(t, i_0)$ (respectively $P_I^j(t, i_0)$, $P_R^j(t, i_0)$) the probability that node j was in state S (resp. in state I and R) at time t provided the patient zero was node i_0 . With the use of Bayes rule, the probability that node i is the patient zero given the observed states is proportional to the joint probability of observed states given the patient zero, $P(i|\mathcal{O}) \sim P(\mathcal{O}|i)$. We can also define an energy-like function of every node $E(i) \equiv -\log P(\mathcal{O}|i)$, nodes with lower energy are then more likely to be the infection source. If one were able to compute $P(\mathcal{O}|i)$ exactly this would be an optimal inference scheme. However, there is no tractable way to compute exactly the joint probability of the observations, hence we approximate it using a mean-field-type approach as a product of the marginal probabilities provided by the dynamic message-passing

$$P(\mathcal{O}|i) \simeq \prod_{\substack{k \in \mathcal{O} \\ q^k(t_0)=S}} P_S^k(t_0, i) \prod_{\substack{l \in \mathcal{O} \\ q^l(t_0)=I}} P_I^l(t_0, i) \prod_{\substack{m \in \mathcal{O} \\ q^m(t_0)=R}} P_R^m(t_0, i). \quad (1)$$

Finally to estimate the right value of time t_0 we need to compute energy $E(i)$ for different possible values of t_0 and choose the minimum one.

Dynamic message-passing algorithm: Let us explain the dynamic message-passing equations for the SIR model. The proof that these equations are exact on trees and connections to some related existing results, namely [13–16, 19–22], are discussed in appendix B. We first define the message $P_S^{i \rightarrow j}(t)$ as the probability for node i to be in the state S at time t in the cavity graph in which node j has been removed. The quantity $\theta^{k \rightarrow i}(t)$ is the probability for node k not to pass the infection signal to node i up to time t , and $\phi^{k \rightarrow i}(t)$ is the probability for node k to be in the state I and not to pass the infection to node i up to time t . The initial condition are $\theta^{k \rightarrow i}(0) = 1$, and $\phi^{k \rightarrow i}(0) = \delta_{q_k(0), I}$. For more precise definitions see appendix A. These messages satisfy the recursion rules:

$$P_S^{i \rightarrow j}(t+1) = P_S^i(0) \prod_{k \in \partial i \setminus j} \theta^{k \rightarrow i}(t+1), \quad (2)$$

$$\theta^{k \rightarrow i}(t+1) - \theta^{k \rightarrow i}(t) = -\lambda_{ki} \phi^{k \rightarrow i}(t), \quad (3)$$

$$\phi^{k \rightarrow i}(t) = (1 - \lambda_{ki})(1 - \mu_k) \phi^{k \rightarrow i}(t-1) - [P_S^{k \rightarrow i}(t) - P_S^k(t-1)]. \quad (4)$$

Here $\partial i \setminus j$ means the set of nodes neighboring node i , excluding j . The marginal probabilities that node i is in a given state at time t are then given as

$$P_S^i(t+1) = P_S^i(0) \prod_{k \in \partial i} \theta^{k \rightarrow i}(t+1), \quad (5)$$

$$P_R^i(t+1) = P_R^i(t) + \mu_i P_I^i(t), \quad (6)$$

$$P_I^i(t+1) = 1 - P_S^i(t+1) - P_R^i(t+1). \quad (7)$$

Hence the algorithmic complexity for computing the energy $E(i)$ of a given vertex i (and therefore the probability that it is the epidemic origin) is $O(t_0 N c)$, where c is the average degree of the graph.

Performance of inference algorithms: We first test our algorithm on random regular graphs, i.e. random graphs where every node has degree c . In all the examples we consider uniform transmission and recovery probabilities $\lambda_{ij} = \lambda$ and $\mu_i = \mu$.

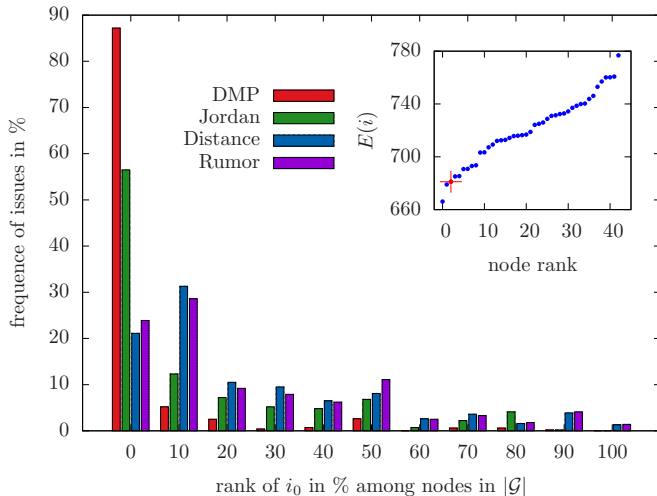


FIG. 1. (color online) A test of inference of the epidemic origin on random regular graphs of degree $c = 4$, size $N = 1000$. Inset: An epidemic is generated with recovery probability $\mu = 1$, transmission probability $\lambda = 0.6$, a snapshot of all the nodes is taken at time $t_0 = 8$ (in this figure we assume we know the value of the time t_0), 242 nodes are observed to be in the I or R state. The dynamic message-passing is used to compute the energy of every node. This energy is finite for 43 nodes; it is plotted as a function of their rank r . The true patient zero is marked by a red cross, and its rank is $r(i_0) = 2$ in this case. The main figure: An epidemic generated with $\mu = 1$, $\lambda = 0.5$, $t_0 = 5$. Histogram (over 1000 random instances) of the normalized rank (i.e. the rank divided by the number of R or I nodes in the snapshot) of the true patient zero is plotted for the dynamic message-passing inference, as well as for the distance, rumor and Jordan centrality measures.

In the first example, inset of Fig. 1, we plot the energy $E(i)$ resulting from the dynamic message-passing of the nodes for which the probability of being the epidemic origin is finite according to our algorithm, we order the nodes according to the energy value. The true epidemic origin is marked with a red cross. We define the rank of candidates for the epidemic origin to be its position in this ranking (the lowest energy node having rank 0). In the main part of Fig. 1 we plot the histogram of normalized ranks (i.e. the rank divided by the total number of nodes that were observed as recovered or infected) of the true epidemic origin as obtained from our DMP inference algorithm, compared to the rankings obtained by distance, rumor and Jordan centralities. We see that the DMP inference algorithm considerably outperforms the three centrality measures, with a comparable computational cost.

In Fig. 2 we present the average normalized rank of the true epidemic origin for random regular graphs for the

whole range of the transmission probability λ , for different values of the recovery probability μ , and a snapshot of all the nodes at time t_0 . We see that DMP inference outperforms the centrality measures, except in case (c) in a range of $0.3 < \lambda < 0.58$ where Jordan center is a better estimation. In other cases, however, Jordan centrality is less performant. Note that for $\mu < 1$ our implementation of Jordan centrality does not distinguish between recovered and infected nodes, which partly explains its very bad performance in that case. We have also computed the results of the rumor centrality measure, according to our results it has never been systematically better than distance centrality.

It is important to note that for some range of parameters the average normalized rank of the true epidemic origin is not so close to zero (note that value $1/2$ of the normalized rank corresponds to a random guess of patient zero among all the infected or recovered nodes). The problem of estimating the epidemic origin with a good precision is very hard in these regions. In some cases the information about the epidemic origin was lost during the spreading process: for instance for $\lambda > \lambda_c = \mu/(c-1)$ [23] the epidemic percolates at large times $t_0 \gg \log_c N$. Then the information about the epidemic origin is lost. On the other hand for $t_0 < \log_c N$, Fig. 2 (b), the epidemic is confined to a tree network and in this case the inference of the origin is easier. In Fig. 2 we mostly focused on the intermediate case $t_0 \approx \log_c N$. In our opinion the systematic comparison presented here is a good test-bed for comparing and improving algorithms.

So far we have tested only cases where the states of all the nodes are observed in the snapshot. In practical situations it is more likely that only a fraction of nodes is observed. We have tested our algorithm in the case where a fraction ξ of nodes is not observed. Fig. 3 (a) shows the average rank of the true epidemic origin. We chose parameters for which our algorithm compared the worse to the Jordan and distance centralities, generalized to the case of incomplete snapshot as described in appendix B. The result of our test in Fig. 3 shows that with incomplete snapshots the DMP inference algorithm is outperforming both centralities, even in the case where for complete snapshots the Jordan centrality was better. Such a robustness is a very useful property.

In the example depicted in Fig. 3 (b), we took the network of the U.S. East-Coast power grid which contains $N = 4941$ nodes with a maximum degree 19 [24]. We see that the DMP estimator gives better prediction for all range of λ .

Finally note that in our numerical tests with DMP we assumed so far the knowledge of the observation time t_0 . To estimate the spreading time we use the following procedure: given a snapshot of epidemic spread, first use one of the distance centrality algorithm in order to select an estimation of the epidemic origin i_0^* . This estimation does not need to be very precise. Then plot the values of the energy $E(i_0^*)$ as a function of time, the minimum of this function is maximizing the probability that i_0^* was

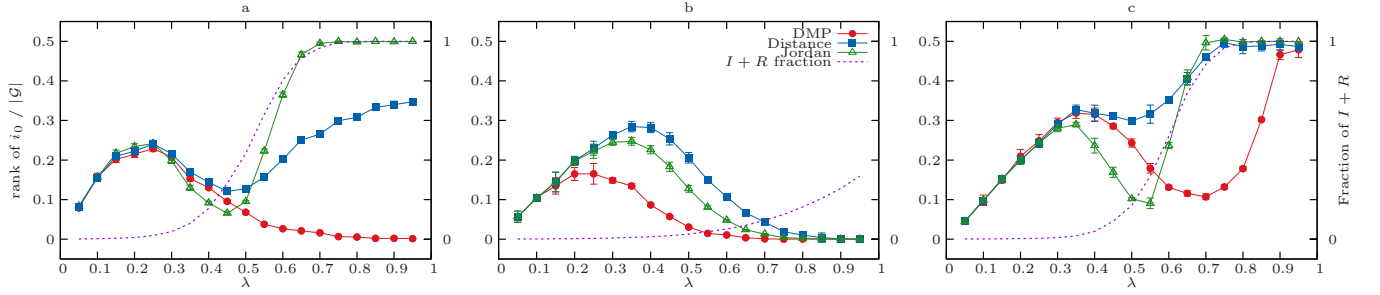


FIG. 2. (color online) Average rank of the true epidemic origin on random regular graphs of size $N = 1000$ with degree $c = 4$. Each data point is averaged over 1000 instances. The snapshot time t_0 and recovery probability μ are from the left: (a) $t_0 = 10$, $\mu = 0.5$, (b) $t_0 = 5$, $\mu = 1$ and (c) $t_0 = 10$, $\mu = 1$. DMP estimator is given by red circles, Jordan centrality estimator is in green triangles, distance centrality estimator is in blue boxes. In a dotted line we plot the average fraction of nodes that were infected or recovered in the snapshot, $|\mathcal{G}|/N$, we use this number to normalize the ranks of the epidemic origin.

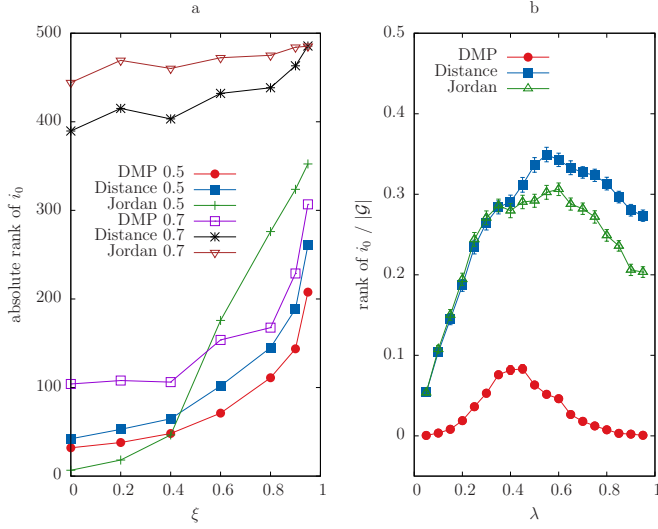


FIG. 3. (color online) Left: (a) Data for a random contact network of size $N = 1000$, degree $c = 4$. Recovery probability $\mu = 1$, transmission probability $\lambda = 0.5$ and $\lambda = 0.7$, only the state of a fraction $1 - \xi$ of nodes is observed at time $t_0 = 10$. Rank (averaged over 1000 instances) of the true epidemic origin obtained with our DMP inference algorithm is compared to the distance and Jordan centralities. Right: (b) Normalized rank (averaged over 1000 instances) of the true epidemic origin for epidemic spreading with $\mu = 0.5$, and all nodes observed at time $t_0 = 10$, on the networks of the U.S. East-Coast power grid. DMP inference is significantly better than inference based on distance and Jordan centralities.

the epidemic origin and lies generally very close to the true value of t_0 .

Our algorithm is based on an approximation to the Bayes optimal inference, and therefore it is not optimal. There are two possible sources of sub-optimality. The first is the fact that the message passing equations may lead to errors on loopy graphs. The second is the mean-field-like approximation (1) of the joint probability distribution. We have observed that taking into account the two-point correlation in this approximation does not lead to any improvement in our results. It would be interesting to search for a better approximations of the likelihood on a general graph.

Conclusion: The solution of dynamics of the SIR model in terms of message-passing equations allowed us to develop an efficient algorithm for detection of the epidemic origin. Compared to existing algorithms, it generically (except for a narrow range of parameters) provides an improved estimate for the source of infectious outbreak and its performance is robust in the case where one has access to the status of only a fraction of the nodes in the network.

Acknowledgments: This work has been supported in part by the EC Grant STAMINA, No. 265496, and by the Grant DySpaN of Triangle de la Physique.

-
- [1] J. Murray, *Mathematical biology* (Springer-Verlag (Berlin and New York), 1989).
 - [2] H. W. Hethcote, *SIAM Rev.* **42**, 599653 (2000).
 - [3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, *Physics Reports* **424**, 175308 (2006).
 - [4] D. Shah and T. Zaman, in *SIGMETRICS'10, Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems* (2010) pp. 203–214.
 - [5] C. H. Comin and L. da Fontoura Costa, *Phys. Rev. E* **84**, 056105 (2011).
 - [6] D. Shah and T. Zaman, *IEEE Trans. Inform. Theory* **57**, 5163 (2011).
 - [7] K. Zhu and L. Ying, “Information source detection in the sir model: A sample path based approach,” (2012), arXiv:1206.5421.
 - [8] B. A. Prakash, J. Vreeken, and C. Faloutsos, in *ICDM'12; Proceedings of the IEEE International Conference on Data Mining* (2012).

- [9] V. Fioriti and M. Chinnici, “Predicting the sources of an outbreak with a spectral technique,” (2012), arXiv:1211.2333 [math-ph].
- [10] P. C. Pinto, P. Thiran, and M. Vetterli, Phys. Rev. Lett. **109**, 068702 (2012).
- [11] C. W. T. Wenxiang Dong, Wenyi Zhang, “Rooting out the rumor culprit from suspects,” (2013), arXiv:1301.6312 [cs.SI].
- [12] W. O. Kermack and A. G. McKendrick, Proc. R. Soc. Lond. A **115**, 700 (1927).
- [13] B. Karrer and M. E. J. Newman, Physical Review E **82**, 016101 (2010).
- [14] E. Volz, J. Math. Biol. **56**, 293310 (2008).
- [15] J. C. Miller, J. Math. Biol. **62**, 349 (2011).
- [16] H. Ohta and S. Sasa, EPL (Europhysics Letters) **90**, 27008 (2010).
- [17] J. Yedidia, W. Freeman, and Y. Weiss, in *Exploring Artificial Intelligence in the New Millennium* (Science & Technology Books, 2003) pp. 239–236.
- [18] M. Mézard and A. Montanari, *Physics, Information, Computation* (Oxford Press, Oxford, 2009).
- [19] E. Volz and L. A. Meyers, Proc. R. Soc. B **274**, 29252933 (2007).
- [20] E. Volz and L. A. Meyers, J. R. Soc. Interface **6**, 233 (2009).
- [21] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. V. den Broeck, and P. Vanhems, BMC Medicine **9**, 87 (2011).
- [22] Y. Kanoria and A. Montanari, Annals of Applied Probability **21**, 5 (2011).
- [23] M. E. J. Newman, Phys. Rev. E **66**, (2002) **66**, 016128 (2002).
- [24] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).

Appendix A: Dynamic message-passing for SIR model

We present here a proof that the probabilities of being susceptible/infected/recovered at a given time t as provided by the dynamic message-passing (DMP) equations (2-7) from the main part of this paper are exact for all initial conditions and every realization of the transmission and recovery probabilities λ_{ij} and μ_i if the graph of contacts is a tree. Before giving the proof we start with a couple of remarks explaining relation to existing works.

1. General remarks about DMP

It should be noted that equations equivalent to (2-7) were first derived in [13]. The authors of [13] treated a more general SIR model where the transmission and recovery probability depends on the time when the node in question was infected. For this more general case an easily tractable form (i.e. the probabilities at time t give probabilities at time $t + 1$ via a set of simple closed equation) of the DMP is not known. The equations in [13] were instead written in a convolutional form that is rather complicated for numerical resolution. The authors noticed that when the probabilities of recovery and transmission are constant then the equations simplify, but did not write a version of the equations that is applicable on a given graph for a given initial condition (actually they only wrote equations averaged over a set of initial conditions). Hence we find it useful to provide the derivation of the DMP on a single graph in their simple iterative form.

For the purpose of this paper we use the DMP on a single instance of the contact network for a given initial condition. However, if an ensemble of initial conditions is given as well as an ensemble of random graphs with a given probability distribution then one can write differential equations for the fraction of nodes that are susceptible/infected/recovered at a given time. These equations were first derived by [14] and appeared also in [13] and [15]. One should not confuse these averaged DMP equations with the “naive” mean field equations that are often written for the SIR model under the assumption of perfect mixing, see e.g. [3]. Whereas the naive mean field equations provide only a very crude approximation for the real probabilities, the equations of [14, 15] are exact in the thermodynamic limit, $N \rightarrow \infty$, as long as in the random graph ensemble a random node does not belong (with a high probability) to a loop of constant (in N) length.

It is interesting to realize that the present DMP equations are applicable also for contact networks that are changing in time. The generalization is straightforward, one only needs to encode the dynamics of the network into time-changing transmission probabilities $\lambda_{ij}(t)$ and use the equations (2-7). The SIR model on dynamically changing networks has been already studied using the graph-averaged version of the DMP equations in [19, 20]. We anticipate that the DMP equations on a single graph will also be useful for studies where specific experimental data about the changing network, such as those of [21], can be used.

Further we want the reader to note the similarities and more importantly differences between DMP and BP[17]. The common point for both DMP and BP is that they are both exact if the underlying network is a tree. The crucial difference is that BP is derived from a stationary Boltzmann-like probability distribution and only the fixed point of the BP equations has a physical meaning. Whereas in the DMP equations (presented here for the SIR model) every step of the iterations corresponds to the physical time in the underlying dynamical process. Note, however, that DMP can be derived from a “dynamic” belief propagation where variables in the corresponding graphical model are the whole trajectories of a given node, see e.g. [22]. Here we will present a more straightforward derivation.

2. Derivation of DMP for the SIR model

Here we present the derivation of equations (2-7) for tree contact networks. We define $P_S^i(t)$, $P_I^i(t)$ and $P_R^i(t)$ as marginal probabilities that $q_i(t) = S$, $q_i(t) = I$ and $q_i(t) = R$. These marginals sum to one and thus

$$P_I^i(t+1) = 1 - P_S^i(t+1) - P_R^i(t+1). \quad (\text{A1})$$

Since the recovery process from state I to state R is independent of neighbors, we have

$$P_R^i(t+1) = P_R^i(t) + \mu_i P_I^i(t). \quad (\text{A2})$$

The epidemic process on a graph can be interpreted as the propagation of infection signals from infected to susceptible nodes. The infection signal $d^{i \rightarrow j}(t)$ is defined as a random variable which is equal to one with probability $\delta_{q_i(t-1), I} \lambda_{ij}$, and equal to zero otherwise. Consider an auxiliary dynamics D_j where the node j receives infection signals, but ignores them and thus is fixed to the S state at all times. Since the infection cannot propagate through the site j in this dynamic setting, different graph branches rooted at node j become independent if the underlying

graph is a tree. This is the natural generalization of the cavity method used for deriving BP (see [18]) to dynamical processes. Notice that the auxiliary dynamics D_j is identical to the original dynamics D for all times such that $q_j(t) = S$. We also define an auxiliary dynamics D_{ij} in which the state of a pair of neighboring sites i and j is always S .

In order to close the system of message-passing equations, we write the remaining update rules in terms of three kinds of cavity messages, defined as follows:

- $\theta^{k \rightarrow i}(t)$ is the probability that the infection signal has not been passed from the node k to the node i up to time t in the dynamics D_i :

$$\theta^{k \rightarrow i}(t) = \text{Prob}^{D_i} \left(\sum_{t'=0}^t d^{k \rightarrow i}(t') = 0 \right); \quad (\text{A3})$$

- $\phi^{k \rightarrow i}(t)$ is the probability that the infection signal has not been passed from the node k to the node i up to time t in the dynamics D_i and that k is in the state I at time t :

$$\phi^{k \rightarrow i}(t) = \text{Prob}^{D_i} \left(\sum_{t'=0}^t d^{k \rightarrow i}(t') = 0, q_k(t) = I \right); \quad (\text{A4})$$

- $P_S^{k \rightarrow i}(t)$ is the probability in the dynamics D_i that k is in the state S at time t :

$$P_S^{k \rightarrow i}(t) = \text{Prob}^{D_i} (q_k(t) = S). \quad (\text{A5})$$

In what follows, we prove that

$$P_S^{i \rightarrow j}(t+1) = P_S^i(0) \prod_{k \in \partial i \setminus j} \theta^{k \rightarrow i}(t+1), \quad (\text{A6})$$

where $\partial i \setminus j$ means the set of neighbors of i excluding j . Indeed, by definition

$$P_S^{i \rightarrow j}(t+1) = \text{Prob}^{D_j} (q_i(t+1) = S) = P_S^i(0) \text{Prob}^{D_j} \left(\sum_{k \in \partial i \setminus j} \sum_{t'=0}^{t+1} d^{k \rightarrow i}(t') \right). \quad (\text{A7})$$

Since the auxiliary dynamics D_{ij} coincides with dynamics D_j as long as the node i is in the S state, we can write

$$P_S^{i \rightarrow j}(t+1) = P_S^i(0) \text{Prob}^{D_{ij}} \left(\sum_{k \in \partial i \setminus j} \sum_{t'=0}^{t+1} d^{k \rightarrow i}(t') \right). \quad (\text{A8})$$

Since different branches of the graph containing nodes $k \in \partial i \setminus j$ are connected only through the node i , they are independent of each other, hence

$$P_S^{i \rightarrow j}(t+1) = P_S^i(0) \prod_{k \in \partial i \setminus j} \text{Prob}^{D_{ij}} \left(\sum_{t'=0}^{t+1} d^{k \rightarrow i}(t') \right), \quad (\text{A9})$$

Moreover, for the nodes $k \in \partial i \setminus j$, the dynamics D_{ij} is equivalent to the dynamics D_i , so we can replace D_{ij} by D_i in the last expression and hence, using the definition (A3), we obtain equation (A6).

We complete the updating rules by writing the equations for $\theta^{k \rightarrow i}(t)$ and $\phi^{k \rightarrow i}(t)$. The only way $\theta^{k \rightarrow i}(t)$ can decrease is by actually transmitting the infection signal from k to i , and this happens with rate λ_{ki} proportionally to the probability that the site k was infected, so we have

$$\theta^{k \rightarrow i}(t+1) - \theta^{k \rightarrow i}(t) = -\lambda_{ki} \phi^{k \rightarrow i}(t). \quad (\text{A10})$$

The change for $\phi^{k \rightarrow i}(t)$ at each time step comes from three different possibilities: either the node k actually sends the infection signal to i (with probability λ_{ki}), either it recovers (with probability μ_k), or it switches to I at this time step, being previously in the S state (this happens with probability $S^{i \rightarrow j}(t-1) - S^{i \rightarrow j}(t)$):

$$\phi^{k \rightarrow i}(t) - \phi^{k \rightarrow i}(t-1) = -\lambda_{ki} \phi^{k \rightarrow i}(t-1) - \mu_k \phi^{k \rightarrow i}(t-1) + \lambda_{ki} \mu_k \phi^{k \rightarrow i}(t-1) + S^{k \rightarrow i}(t-1) - S^{k \rightarrow i}(t). \quad (\text{A11})$$

The third compensation term on the right hand side of the previous equation comes to avoid counting twice the situation when the node k transmits the infection and recovers at the same time step. This completes the update rules for cavity messages. These equations can be iterated in time starting from initial conditions for cavity messages:

$$\theta^{i \rightarrow j}(0) = 1, \quad (\text{A12})$$

$$\phi^{i \rightarrow j}(0) = \delta_{q_i(0), I}. \quad (\text{A13})$$

The marginal probability in the original dynamics D is obtained by including all the neighbors $k \in \partial i$ in eq. (A6):

$$P_S^i(t+1) = P_S^i(0) \prod_{k \in \partial i} \theta^{k \rightarrow i}(t+1). \quad (\text{A14})$$

The closed set of equations (A1,A2,A6,A10,A11,A14) , together with the initial conditions (A12-A13), give the exact values of marginal probabilities $P_S^i(t)$, $P_I^i(t)$ and $P_R^i(t)$ on a tree graph.

Appendix B: The centrality algorithms for incomplete snapshots

In the case where the state of all the nodes is known at time t_0 , the centrality algorithms work on a connected component \mathcal{G} of infected and recovered nodes. In practice the information is available only for a fraction $1 - \xi$ of nodes in the graph G . The snapshot $\mathcal{O}(t_0)$ can then be thought of as a configuration of $(1 - \xi)N$ nodes in the states S , I , R (nodes for which we have the information), and of ξN randomly located nodes in the unknown state X . Now the infected and recovered nodes in general do not form a connected component and are located in several disconnected components, separated by the nodes in the unknown states X . Nevertheless, it is clear that not all the X -nodes have to be checked as possible candidates to be the actual source of infection. If the cluster of nodes in the X state is surrounded only by the S -nodes, this cluster is clearly in the S state itself. Other X -nodes in principle are susceptible to be the infection source and thus need to be checked.

We propose the following generalization of centrality algorithms for the $\xi \neq 0$ case. First we construct a connected component composed of all the nodes in the I and R states and clusters of X nodes which are not completely encircled by S -nodes. This gives a connected component of I , R and X nodes attached together. Since now we have a connected component \mathcal{G} , we can run centrality algorithms on it in a usual way. For $\xi = 0$ the connected component constructed in this way coincides with a connected component composed of infected and recovered component.

Below (figure 4) we compare the distributions of ranks for DMP and Jordan estimators for different ξ ($\xi = 0$, $\xi = 0.5$ and $\xi = 0.9$) in the special case of 'deterministic' recovery $\mu = 1$ for $\lambda = 0.5$ and $\lambda = 0.7$. The results are presented for a regular random graph composed of $N = 1000$ nodes with connectivity $c = 4$, and we take $t_0 = 10$. The plot shows how often the rank of the actual epidemic origin i_0 is within the value of the corresponding bin (0% means exact reconstruction). According to the histogram, in 60% of cases we manage to locate the true infection source within 10% of relevant nodes (those situated in \mathcal{G}) for $\xi = 0$. This number falls to 40% for $\xi = 0.9$, when the states of only 10% of nodes in the network are known.

We see that although for $\xi = 0$ the rank distribution based on the Jordan centrality estimator gives better results (in the case $\lambda = 0.5$), it is no longer efficient when the number of unknown nodes gets larger (for all $\xi > 0.4$). The dependence on ξ for the case $\lambda = 0.7$ follows the same patterns.

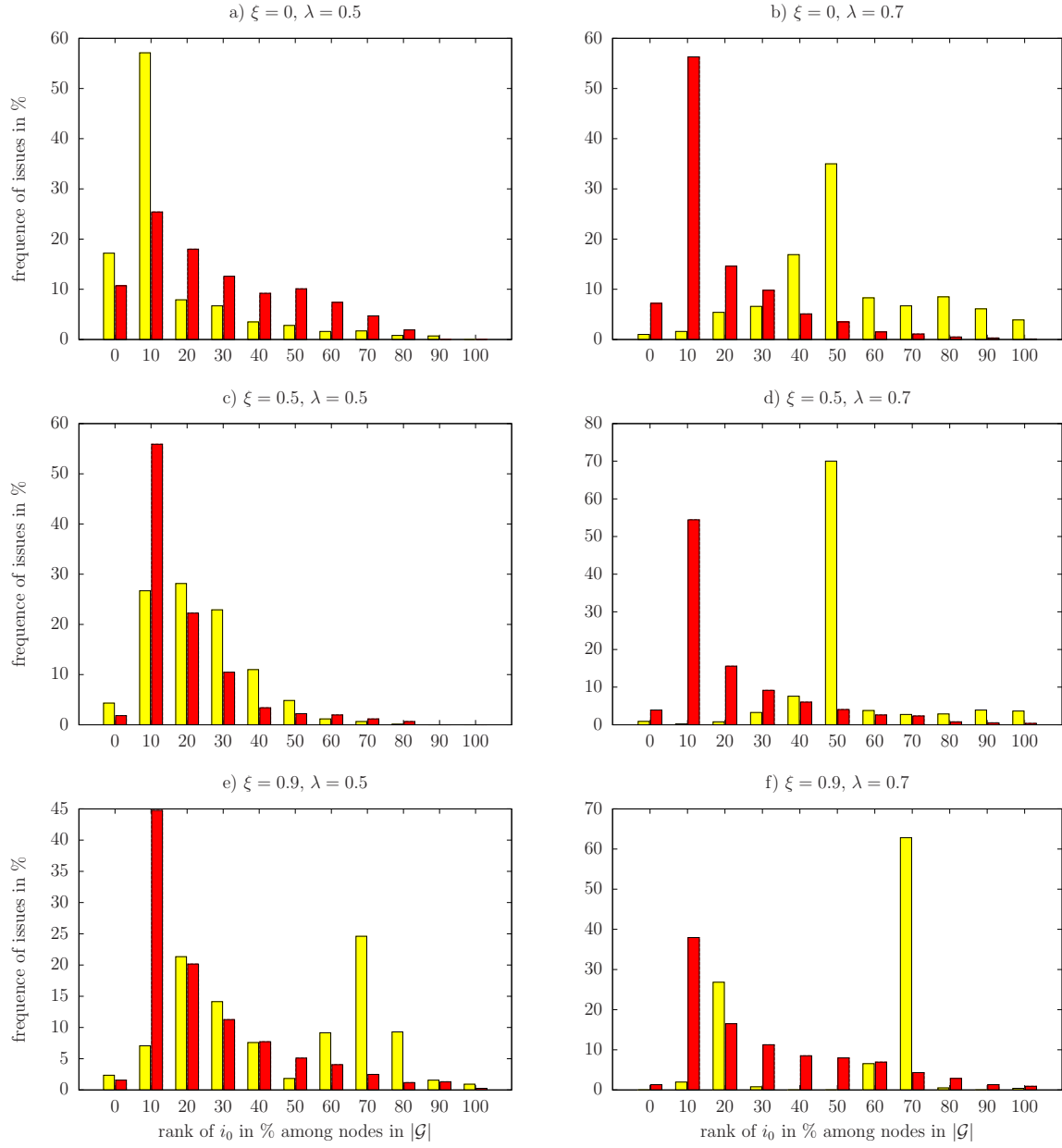


FIG. 4. (color online) Distribution of inferred rank of the epidemic origin measured over the graph \mathcal{G} for Jordan centrality estimator (yellow) and DMP estimator (red) on regular random graphs: a) $\xi = 0, \lambda = 0.5$, b) $\xi = 0, \lambda = 0.7$, c) $\xi = 0.5, \lambda = 0.5$, d) $\xi = 0.5, \lambda = 0.7$, e) $\xi = 0.9, \lambda = 0.5$, f) $\xi = 0.9, \lambda = 0.7$. The average is performed over 500 instances.